Stat 511 HW#4 Spring 2009

1. Below are some slightly fake data reconstructed from information in the article "Process Optimization and Characterization of Poloxamer Solid Dispersions of a Poorly Water-soluble Drug" by Shah, Amin, Parikh and Parikh. Given are

$y =$ time required for 90% drug dissolution (min)

$x_1 =$ (coded) temperature to which the melt-drug mixture cooled

$x_2 =$ (coded) drug-to-polymer ratio

for $n = 11$ solid dispersions of PXM-RXB prepared to study the effects of the process variables on drug dissolution characteristics.

x_1	x_2	y
–1	–1	62.9
–1	0	21.5
–1	1	19.3
0	–1	64.1
0	0	32.9
0	1	30.9
1	–1	93.4
1	0	64.6
1	1	58.6
–.5	.33	26.6
.5	–.33	51.3

Consider the two regression models

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \qquad (1)$$

and

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} + \varepsilon_i \qquad (2)$$

Use R to do the following.

a) Fit the two models using lm( ) and normal plot standardized residuals.

b) In the model (2) test $\mathrm{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0$. Report a *p*-value. Does quadratic curvature in response (as a function of the x's) appear to be statistically detectable?

c) Use multivariate calculus on the fitted quadratic equation and find the location (x_1, x_2) of a stationary point of the fitted quadratic surface. Is this the location of an absolute maximum or absolute minimum of the fitted response?

d) For both fitted models, find the (x_1, x_2) location of smallest estimated mean time to 90% dissolution. For these models and locations, use R matrix calculations to find 90% two-sided confidence limits for the mean responses there. Then find 90% two-sided prediction limits for a single new response from these sets of conditions under these models.

2. (Testing "Lack of Fit" … See Section 6.6 of Christensen) Suppose that in the usual linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\mathbf{X}$ is of full rank (k). Suppose further that there are $m < n$ distinct rows in $\mathbf{X}$ and that $m > k$. One can then make up a "cell means" model for $\mathbf{Y}$ (where observations having the same corresponding row in $\mathbf{X}$ are given the same mean response) say

$$\mathbf{Y} = \mathbf{X}^{*}\boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

This model puts no restrictions on the means of the observations except that those with identical corresponding rows of $\mathbf{X}$ are equal. It is the case that $C(\mathbf{X}) \subset C(\mathbf{X}^{*})$ and it thus makes sense to test the hypothesis $\mathrm{H}_0\text{:E}\mathbf{Y} \in C(\mathbf{X})$ in the cell means model. This can be done using

$$F = \frac{\mathbf{Y}'\left(\mathbf{P}_{\mathbf{X}^{*}} - \mathbf{P}_{\mathbf{X}}\right)\mathbf{Y}/(m-k)}{\mathbf{Y}'\left(\mathbf{I} - \mathbf{P}_{\mathbf{X}^{*}}\right)\mathbf{Y}/(n-m)}$$

and this is usually known as testing for "lack of fit."

Consider again the situation in Problem 1. Suppose that in fact there were two additional observations at $(x_1, x_2) = (0,0)$, namely $y = 30$ and $y = 36$. Use `R` and matrix calculations to find a *p*-value for testing lack of fit to the quadratic regression model (2) in Problem 1.

3. Below is a small table of fake 2-way factorial data. Enter them into `R` in three vectors of length $n = 13$. Call these vectors "`y`", "`A`", and "`B`".

	Level 1 of B	Level 2 of B	Level 3 of B
Level 1 of A	12	13,14	20
Level 2 of A	8	10	6,7,8
Level 3 of A	10,11	13	7

a) Create and print out an `R` data frame using the commands

```
> d<-data.frame(y,A,B)
> d
```

b) Turn the numerical variables `A and B` into variables that `R` will recognize as levels of qualitative factors by issuing the commands

```
> d$A<-as.factor(d$A)
> d$B<-as.factor(d$B)
```

Then compute and print out the cell means by typing

```
> means<-tapply(d$y,list(d$A,d$B),mean)
> means
```

You may find out more about the function `tapply` by typing

```
> ?tapply
```

c) Make a crude interaction plot by doing the following. First type

```
> x.axis<-unique(d$B)
```

to set up horizontal plotting positions for the sample means. Then make a "matrix plot" with lines connecting points by issuing the commands

```
> matplot(c(1,3),c(5,25),type="n",xlab="B",ylab="Mean
Response",main="y")
> matlines(x.axis,means,type="b")
```

The first of these commands sets up the axes and makes a dummy plot with invisible points "plotted" at $(1,5)$ and $(3,25)$. The second puts the lines and identifying `A` levels (as plotting symbols) on the plot.

d) Set the default for the restriction used to create a full rank model matrix, run the linear models routine and find both sets of "Type I" sums of squares by issuing the following commands

```
> options(contrasts=c("contr.sum","contr.sum"))
> lm.out1<-lm(y~A*B,data=d)
> summary.aov(lm.out1,ssType=1)
> lm.out2<-lm(y~B*A,data=d)
> summary.aov(lm.out2,ssType=1)
```

See if anything changes if you ask R to compute "Type III" sums of squares by issuing the command

```
> summary.aov(lm.out1,ssType=3)
```

(In the past `R` has failed to respond to the request for Type III sums of squares without warning you that it is going to fail to do so.)

e) Start over with this problem, doing the calculations "from scratch" using your basic linear models knowledge and matrix calculations in R. Compute all of Type I, Type II and Type III sums of squares here, using the sum restriction in the first two cases (and the order of factors A,B). Then compute Type I and Type II sums of squares using the SAS baseline restriction.

f) Now suppose that by some misfortune, the observation from the $(3,1)$ cell of this complete 3×3 factorial somehow gets lost and one has only $n=11$ observations from $k=8$ cells (and thus "incomplete factorial" data). Test the hypothesis that at least for the cells where one has data, there are no interactions, i.e. $\mathrm{E}\,\mathbf{Y} \in C\left(\left(\mathbf{1} \mid \mathbf{X}_{\alpha^*} \mid \mathbf{X}_{\beta^*}\right)\right)$. (Note that this matrix $\left(\mathbf{1} \mid \mathbf{X}_{\alpha^*} \mid \mathbf{X}_{\beta^*}\right)$ should be of full rank.)

g) In the incomplete factorial context of part f), the function $\mu + \alpha_1^* + \beta_3^*$ is estimable. What is the OLS estimate for it? (Note that this is the mean response for the missing cell only if the same no-interaction model used to describe the 8 cells extends to the 9th. This is the kind of assumption one makes in regression analysis when using a fitted prediction equation to estimate a mean response at a set of conditions not in one's original data set. It might well be argued, however, that the link between observed and unobserved conditions is intuitively stronger with quantitative factors than with qualitative factors.)

4. (Refer to 2008 HW#5, 2004 HW#6 2004 and 2003 HW #7 (and the first part of 2003 HW#8) and the associated keys for patterns for the `R` code needed to do the following problem.)

The book *Statistical Tools for Nonlinear Regression* by Huet, Bouvier, Poursat, and Jolivet a data set a number of nice nonlinear regression data sets in it. One is below. Provided are measured optical density, y, (indicative of antibody level) for various dilutions of serum samples taken from the same cow in May and then in June in a so-called ELISA analysis of the cow's serum.

Dilution, d		1/30	1/90	1/270	1/810	1/2430	1/7290	1/21870	1/65610
Optical Density, y	May	1.909 1.956	1.856 1.876	1.838 1.841	1.579 1.584	1.057 1.072	0.566 0.561	0.225 0.229	0.072 0.114
	June	1.886 1.880	1.853 1.870	1.747 1.772	1.424 1.406	0.781 0.759	0.377 0.376	0.153 0.138	0.053 0.058

Let

$$x = -\log_{10} d$$

For data from a single month, a standard model for data like these is

$$y_i = \frac{\theta_1}{1 + \exp\left(\theta_2\left(x_i - \theta_3\right)\right)} + \varepsilon_i$$

Note that θ_1 is the limiting mean for small x, 0 is the limiting mean for large x, the value θ_3 is x at which the mean is $\theta_1/2$, and the mean is $\theta_1/4$ at $x = \theta_3 + \ln 3/\theta_2$. (These facts can be used to develop "eye estimates" of the parameters $\theta_1, \theta_2,$ and θ_3 from plots of y vs x.)

a) Plot y vs x and make "eye estimates" of the model parameters for both months. Use these as starting values and employ the `nls` package in `R` to separately fit both models. Then replot the data along with 2 fitted mean curves (one for each month) on the same set of axes. Do the two mean functions "look" to be similar? If they are not the same, does it seem plausible that one is a right-shifted version of the other? Normal plot the two sets of residuals. On the basis of these, do you see obvious problems with an "iid normal" model for the ε_i in either of the months?

b) Verify for the May data that `nls` outputs $MSE\left(\hat{\mathbf{D}}'\hat{\mathbf{D}}\right)^{-1}$ as an estimated covariance matrix for $\left(\hat{\theta}_{1\text{OLS}},\hat{\theta}_{2\text{OLS}},\hat{\theta}_{3\text{OLS}}\right)'$ and that the standard errors it produces are the square roots of diagonal elements of this matrix. Use this matrix and make approximate 95% confidence limits for the mean optical density at a 1/270 dilution, and then approximate 95% prediction limits for an observed May optical density at that dilution not included in the original data set.

c) The dilution at which May mean optical density reaches 25% of its (small x) limiting value is a function of $\boldsymbol{\theta}$. Find a sensible point estimate of this quantity and an estimated standard deviation for this estimate.

d) Make 95% t intervals for the three May parameters based on the estimated covariance matrix from b). Add the `MASS` package to your `R` environment in order to get access to the function `confint()` and use it to make approximate 95% intervals for $\theta_1,\theta_2,$ and θ_3. How do these intervals compare to the t intervals?

e) Make two different approximately 95% confidence intervals for the May σ, one based on carrying over the linear model result that $SSE/\left(n-k\right)\sim\chi^2_{n-k}$ to the nonlinear model, and the other based on the "profile likelihood" material. Does the function `confint()` provide limits for σ?

f) A 4-parameter model that forces May and June mean functions to simply be shifted versions of each other is

$$y_i=\frac{\theta_1}{1+\exp\left(\theta_2\left(x_i-\left(\theta_3+\theta_4\cdot I\left[\text{case } i \text{ is from June}\right]\right)\right)\right)}+\varepsilon_i$$

Fit this model to the whole data set (both May and June observations). Make a plot of both mean functions (May and June) on the same set of axes and compare to the plot you made in a). Does this model fit nearly as well as the two separate fits in a)? Explain. The quantity 10^{θ_4} is sometimes called the "relative potency" of the serum at the two periods. Make 95% confidence limits for this quantity.

There is material on earlier non-linear regression problems from 2003, 2004, and 2008 homeworks not included here (basically because there are more parameters in this problem than in the 2-parameter set-up used in those earlier homeworks and plotting of loglikelihoods is not feasible). Carefully study those parts of the old HWs. You are responsible for that material.